Big Data Examination

Roll No. - DS5B-2121

Question 2

Q.II Considering the Item_Outlet_Sales as dependent variable in “Big Mart Sales” dataset, determine the accuracy of the model. Split the dataset according to your last digit of roll no. (Example: if your roll no is ending with 0, the ratio will be 70, 30; if your roll no is ending with 1, the ratio will be 71, 29; if your roll no is ending with 2, the ratio will be 72, 28; if your roll no is ending with 3, the ratio will be 73, 27 etc.).

Importing Pyspark Library

In [ ]:
```
pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publ
ic/simple/
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 32 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 69.0 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=281853642
sha256=bf2c426d86882346b14a6fc9d4f4fe8f8f6c664e20345f63052db0e57c752d41
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e92efd1e1d53
34db05f2e83bcef74f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

Import Library and Creating Session

In [ ]:
```
from pyspark.sql import SparkSession
```

In [ ]:
```
session = SparkSession.builder.appName("Big_Mart_Sales_Dataset").master("local").getOrCr
#we reassign value of __name__ (inbuilt variable) to "__main__" and main is used as entr
# else the value of name might be different
```

Reading the Data From CSV

In [ ]:
```
data = session.read.csv("Big Mart Sale.csv", header = True, inferSchema=True)
```

To print top 10 raw in dataset

In [ ]:
```
data.show(10)
```

```
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
```

```
----------+-----------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_
MRP|Outlet_Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|      O
utlet_Type|Item_Outlet_Sales|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|          FDA15|        9.3|         Low Fat|    0.016047301|               Dairy|249.8
092|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|         3735.138|
|          DRC01|       5.92|         Regular|    0.019278216|         Soft Drinks| 48.2
692|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         443.4228|
|          FDN15|       17.5|         Low Fat|    0.016760075|                Meat| 141.
618|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|          2097.27|
|          FDX07|       19.2|         Regular|            0.0|Fruits and Vegeta...| 182.
095|           OUT010|                     1998|       null|              Tier 3|    Gro
cery Store|           732.38|
|          NCD19|       8.93|         Low Fat|            0.0|           Household| 53.8
614|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         994.7052|
|          FDP36|     10.395|         Regular|            0.0|        Baking Goods| 51.4
008|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         556.6088|
|          FDO10|      13.65|         Regular|    0.012741089|         Snack Foods| 57.6
588|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         343.5528|
|          FDP10|       null|         Low Fat|    0.127469857|         Snack Foods|107.7
622|           OUT027|                     1985|     Medium|              Tier 3|Superma
rket Type3|        4022.7636|
|          FDH17|       16.2|         Regular|    0.016687114|        Frozen Foods| 96.9
726|           OUT045|                     2002|       null|              Tier 2|Superma
rket Type1|        1076.5986|
|          FDU28|       19.2|         Regular|     0.09444959|        Frozen Foods|187.8
214|           OUT017|                     2007|       null|              Tier 2|Superma
rket Type1|         4710.535|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
only showing top 10 rows
```

Check Null Values in columns

In [ ]:

```
from pyspark.sql.functions import isnan, when, count, col
```

In [ ]:

```
data.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in data.columns])
```

```
+---------------+-----------+----------------+---------------+---------+--------+-------
----------+-------------------------+-----------+--------------------+-----------+------
-----------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|Item_Type|Item_MRP|Outlet_
Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|Outlet_Type|Item_O
utlet_Sales|
+---------------+-----------+----------------+---------------+---------+--------+-------
----------+-------------------------+-----------+--------------------+-----------+------
-----------+
|              0|       1463|               0|              0|        0|       0|
         0|                        0|       2410|                   0|          0|
          0|
+---------------+-----------+----------------+---------------+---------+--------+-------
----------+-------------------------+-----------+--------------------+-----------+------
```

```
-----------+
```

Importing sql functions

In [ ]:
```
import pyspark.sql.functions as func
```

In [ ]:
```
data.agg(func.percentile_approx("Item_Weight", 0.5).alias("mean")).show()
```

```
+----+
|mean|
+----+
|12.6|
+----+
```

Fill Null Values

In [ ]:
```
# replace 12.6 in place of Null values in Item_weight column because it is mean in this
data = data.na.fill(value=12.6,subset=["Item_Weight"])
```

In [ ]:
```
# return Medium in place of Null values in Outlet_Size Column Because Medium is the medi
data = data.na.fill(value="Medium",subset=["Outlet_Size"])
```

In [ ]:
```
data.show()
```

```
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_
MRP|Outlet_Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|      O
utlet_Type|Item_Outlet_Sales|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|          FDA15|        9.3|         Low Fat|    0.016047301|               Dairy|249.8
092|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|         3735.138|
|          DRC01|       5.92|         Regular|    0.019278216|         Soft Drinks| 48.2
692|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         443.4228|
|          FDN15|       17.5|         Low Fat|    0.016760075|                Meat| 141.
618|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|          2097.27|
|          FDX07|       19.2|         Regular|            0.0|Fruits and Vegeta...| 182.
095|           OUT010|                     1998|     Medium|              Tier 3|    Gro
cery Store|           732.38|
|          NCD19|       8.93|         Low Fat|            0.0|           Household| 53.8
614|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         994.7052|
|          FDP36|     10.395|         Regular|            0.0|        Baking Goods| 51.4
008|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         556.6088|
|          FDO10|      13.65|         Regular|    0.012741089|         Snack Foods| 57.6
588|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         343.5528|
|          FDP10|       12.6|         Low Fat|    0.127469857|         Snack Foods|107.7
622|           OUT027|                     1985|     Medium|              Tier 3|Superma
rket Type3|        4022.7636|
|          FDH17|       16.2|         Regular|    0.016687114|        Frozen Foods| 96.9
```

```
726|           OUT045|                     2002|     Medium|               Tier 2|Superma
rket Type1|        1076.5986|
|          FDU28|       19.2|         Regular|      0.09444959|        Frozen Foods|187.8
214|           OUT017|                     2007|     Medium|               Tier 2|Superma
rket Type1|         4710.535|
|          FDY07|       11.8|         Low Fat|             0.0|Fruits and Vegeta...| 45.5
402|           OUT049|                     1999|     Medium|               Tier 1|Superma
rket Type1|        1516.0266|
|          FDA03|       18.5|         Regular|     0.045463773|               Dairy|144.1
102|           OUT046|                     1997|      Small|               Tier 1|Superma
rket Type1|         2187.153|
|          FDX32|       15.1|         Regular|       0.1000135|Fruits and Vegeta...|145.4
786|           OUT049|                     1999|     Medium|               Tier 1|Superma
rket Type1|        1589.2646|
|          FDS46|       17.6|         Regular|     0.047257328|         Snack Foods|119.6
782|           OUT046|                     1997|      Small|               Tier 1|Superma
rket Type1|        2145.2076|
|          FDF32|      16.35|         Low Fat|       0.0680243|Fruits and Vegeta...|196.4
426|           OUT013|                     1987|       High|               Tier 3|Superma
rket Type1|         1977.426|
|          FDP49|        9.0|         Regular|     0.069088961|           Breakfast| 56.3
614|           OUT046|                     1997|      Small|               Tier 1|Superma
rket Type1|        1547.3192|
|          NCB42|       11.8|         Low Fat|     0.008596051|  Health and Hygiene|115.3
492|           OUT018|                     2009|     Medium|               Tier 3|Superma
rket Type2|        1621.8888|
|          FDP49|        9.0|         Regular|     0.069196376|           Breakfast| 54.3
614|           OUT049|                     1999|     Medium|               Tier 1|Superma
rket Type1|         718.3982|
|          DRI11|       12.6|         Low Fat|     0.034237682|         Hard Drinks|113.2
834|           OUT027|                     1985|     Medium|               Tier 3|Superma
rket Type3|         2303.668|
|          FDU02|      13.35|         Low Fat|      0.10249212|               Dairy|230.5
352|           OUT035|                     2004|      Small|               Tier 2|Superma
rket Type1|        2748.4224|
+---------------+-----------+----------------+----------------+--------------------+-----
---+------------------+------------------------+-----------+---------------------+-------
----------+-----------------+
only showing top 20 rows
```

Exploration Of Dataset

To print all columns name

In [ ]:
```
data.columns
```

Out[ ]:
```
['Item_Identifier',
 'Item_Weight',
 'Item_Fat_Content',
 'Item_Visibility',
 'Item_Type',
 'Item_MRP',
 'Outlet_Identifier',
 'Outlet_Establishment_Year',
 'Outlet_Size',
 'Outlet_Location_Type',
 'Outlet_Type',
 'Item_Outlet_Sales']
```

To count total numbers of raws in dataset

In [ ]:
```
data.count()
```

Out[ ]:

```
8523
```

In [ ]:

```python
data.printSchema()
```

```
root
 |-- Item_Identifier: string (nullable = true)
 |-- Item_Weight: double (nullable = false)
 |-- Item_Fat_Content: string (nullable = true)
 |-- Item_Visibility: double (nullable = true)
 |-- Item_Type: string (nullable = true)
 |-- Item_MRP: double (nullable = true)
 |-- Outlet_Identifier: string (nullable = true)
 |-- Outlet_Establishment_Year: integer (nullable = true)
 |-- Outlet_Size: string (nullable = false)
 |-- Outlet_Location_Type: string (nullable = true)
 |-- Outlet_Type: string (nullable = true)
 |-- Item_Outlet_Sales: double (nullable = true)
```

To know data type of each columns

In [ ]:

```python
data.dtypes
```

Out[ ]:

```
[('Item_Identifier', 'string'),
 ('Item_Weight', 'double'),
 ('Item_Fat_Content', 'string'),
 ('Item_Visibility', 'double'),
 ('Item_Type', 'string'),
 ('Item_MRP', 'double'),
 ('Outlet_Identifier', 'string'),
 ('Outlet_Establishment_Year', 'int'),
 ('Outlet_Size', 'string'),
 ('Outlet_Location_Type', 'string'),
 ('Outlet_Type', 'string'),
 ('Item_Outlet_Sales', 'double')]
```

Data Preprocessing

Here we convert the data into machine readable form

In [ ]:

```python
from pyspark.ml.feature import VectorAssembler, StringIndexer, OneHotEncoder
# It is use for mapping a string column to a index column that will be treated as a cate
# It is feature transformer that combine multiple columns into a single vector column.
# Pyspark ml models takes only one independent variable and one dependent varibale
#but, we have multiple independent variabales, so we use vector assembler to convert the
# of independent variables
```

In [ ]:

```python
# Doing String Indexing
str_index = StringIndexer(inputCols = ['Item_Identifier','Item_Fat_Content','Item_Type',
```

In [ ]:

```python
# One Hot Encoding
one_hot = OneHotEncoder(inputCols =['Item_Identifier1','Item_Fat_Content1','Item_Type1',
```

In [ ]:

```python
# APply Vector Assembler
vector_ass = VectorAssembler(inputCols = ['Item_Weight','Item_Fat_Content2','Item_Visibi
```

Import Linear Regression and Create Model

In [ ]:

```python
from pyspark.ml.regression import LinearRegression
```

```
In [ ]: linear = LinearRegression(featuresCol="allfeatures", labelCol="Item_Outlet_Sales")
```

Create Pipeline for ML Model

```
In [ ]: from pyspark.ml import Pipeline
        mypipeline = Pipeline(stages = [str_index, one_hot, vector_ass, linear])
```

Making Train Test Split

Splitting the Dataset

As my roll no is DS5B-2121 I will be using split as 0.71 and 0.29

```
In [ ]: training, test = data.randomSplit([0.71, 0.29])
```

Model Training

```
In [ ]: lin_reg_model = mypipeline.fit(training)
```

Test Model

```
In [ ]: result = lin_reg_model.transform(test)
```

```
In [ ]: result.show()
```

```
+---------------+-----------+----------------+---------------+------------+--------+-----
------------+-------------------------+-----------+--------------------+-----------------
-+-----------------+----------------+-----------------+----------+------------------+---
-----------------------+------------+---------------------+------------+-----------------
---+-----------------+---------------+------------------+--------------------------+-----
--------+---------------------+-------------+--------------------+------------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|   Item_Type|Item_MRP|Outle
t_Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|       Outlet_Typ
e|Item_Outlet_Sales|Item_Identifier1|Item_Fat_Content1|Item_Type1|Outlet_Identifier1|Out
let_Establishment_Year1|Outlet_Size1|Outlet_Location_Type1|Outlet_Type1|    Item_Identifi
er2|Item_Fat_Content2|     Item_Type2|Outlet_Identifier2|Outlet_Establishment_Year2| Outl
et_Size2|Outlet_Location_Type2| Outlet_Type2|         allfeatures|        prediction|
+---------------+-----------+----------------+---------------+------------+--------+-----
------------+-------------------------+-----------+--------------------+-----------------
-+-----------------+----------------+-----------------+----------+------------------+---
-----------------------+------------+---------------------+------------+-----------------
---+-----------------+---------------+------------------+--------------------------+-----
--------+---------------------+-------------+--------------------+------------------+
|          DRA12|       11.6|         Low Fat|            0.0|Soft Drinks|141.6154|
      OUT045|                     2002|     Medium|              Tier 2|Supermarket Type
1|        3829.0158|           521.0|              0.0|       8.0|                1.0|
                    2.0|         0.0|                  1.0|         0.0| (1543,[521],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|     (9,[1],[1.0])|             (8,[2],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,1,14,21,22...|2225.0090020780335|
|          DRA12|       11.6|         Low Fat|    0.040911824|Soft Drinks|142.3154|
      OUT013|                      1987|       High|               Tier 3|Supermarket Type
1|        2552.6772|          521.0|                0.0|       8.0|                 6.0|
                    6.0|          2.0|                    0.0|          0.0| (1543,[521],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|                (8,[6],[1.0])|
(2,[],[])|         (2,[0],[1.0])|(3,[0],[1.0])|(29,[0,1,5,14,21,...| 2224.685214756014|
|          DRA24|      19.35|          Regular|    0.039895009|Soft Drinks|162.4868|
      OUT013|                      1987|       High|               Tier 3|Supermarket Type
1|        4422.2436|           77.0|                1.0|       8.0|                 6.0|
                    6.0|          2.0|                    0.0|          0.0|  (1543,[77],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|               (8,[6],[1.0])|
(2,[],[])|         (2,[0],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 2597.340924578537|
|          DRA59|       8.27|          Regular|    0.127927931|Soft Drinks|184.8924|
      OUT046|                      1997|      Small|               Tier 1|Supermarket Type
1|        4442.2176|           78.0|                1.0|       8.0|                 2.0|
                    3.0|          1.0|                    2.0|          0.0|  (1543,[78],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[2],[1.0])|                (8,[3],[1.0])|(2,
[1],[1.0])|              (2,[],[])|(3,[0],[1.0])|(29,[0,2,5,14,21,...|2982.1821632476695|
|          DRA59|       12.6|          Regular|    0.127308434|Soft Drinks|186.6924|
      OUT027|                      1985|     Medium|               Tier 3|Supermarket Type
3|        7033.5112|           78.0|                1.0|       8.0|                 4.0|
                    0.0|          0.0|                    0.0|          3.0|  (1543,[78],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[4],[1.0])|                (8,[0],[1.0])|(2,
[0],[1.0])|         (2,[0],[1.0])|    (3,[],[])|(29,[0,2,5,14,21,...| 4331.305955126552|
|          DRB01|       7.39|          Low Fat|    0.082170947|Soft Drinks| 190.953|
      OUT013|                      1987|       High|               Tier 3|Supermarket Type
1|         2466.789|         1271.0|                0.0|       8.0|                 6.0|
                    6.0|          2.0|                    0.0|          0.0|(1543,[1271],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|                (8,[6],[1.0])|
(2,[],[])|         (2,[0],[1.0])|(3,[0],[1.0])|(29,[0,1,5,14,21,...|2966.2617325300225|
|          DRB13|      6.115|          Regular|    0.007055292|Soft Drinks| 188.653|
      OUT049|                      1999|     Medium|               Tier 1|Supermarket Type
1|         3605.307|          522.0|                1.0|       8.0|                 5.0|
                    5.0|          0.0|                    2.0|          0.0| (1543,[522],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[5],[1.0])|                (8,[5],[1.0])|(2,
[0],[1.0])|              (2,[],[])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 2973.506639206491|
|          DRB24|      8.785|          Low Fat|    0.020609218|Soft Drinks|155.1656|
      OUT049|                      1999|     Medium|               Tier 1|Supermarket Type
1|        4016.1056|          906.0|                0.0|       8.0|                 5.0|
                    5.0|          0.0|                    2.0|          0.0| (1543,[906],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[5],[1.0])|                (8,[5],[1.0])|(2,
[0],[1.0])|              (2,[],[])|(3,[0],[1.0])|(29,[0,1,5,14,21,...|2420.7896947929494|
|          DRB25|       12.6|          Low Fat|     0.06912336|Soft Drinks|106.0938|
      OUT027|                      1985|     Medium|               Tier 3|Supermarket Type
3|        2787.0388|          247.0|                0.0|       8.0|                 4.0|
                    0.0|          0.0|                    0.0|          3.0| (1543,[247],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[4],[1.0])|                (8,[0],[1.0])|(2,
[0],[1.0])|         (2,[0],[1.0])|    (3,[],[])|(29,[0,1,5,14,21,...|3046.1042772127885|
|          DRB48|       12.6|          Regular|    0.024733134|Soft Drinks| 40.2822|
      OUT027|                      1985|     Medium|               Tier 3|Supermarket Type
3|        1296.3126|          248.0|                1.0|       8.0|                 4.0|
                    0.0|          0.0|                    0.0|          3.0| (1543,[248],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[4],[1.0])|                (8,[0],[1.0])|(2,
[0],[1.0])|         (2,[0],[1.0])|    (3,[],[])|(29,[0,2,5,14,21,...| 2058.027429105631|
|          DRB48|      16.75|          Regular|    0.024832806|Soft Drinks| 38.7822|
      OUT013|                      1987|       High|               Tier 3|Supermarket Type
1|         667.7974|          248.0|                1.0|       8.0|                 6.0|
                    6.0|          2.0|                    0.0|          0.0| (1543,[248],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|                (8,[6],[1.0])|
(2,[],[])|         (2,[0],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 656.3620458837856|
|          DRC01|       5.92|          Regular|    0.019184026|Soft Drinks| 50.3692|
      OUT013|                      1987|       High|               Tier 3|Supermarket Type
1|         591.2304|          907.0|                1.0|       8.0|                 6.0|
                    6.0|          2.0|                    0.0|          0.0| (1543,[907],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|                (8,[6],[1.0])|
(2,[],[])|         (2,[0],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 806.9476679936189|
|          DRC01|       5.92|          Regular|    0.019238942|Soft Drinks| 49.8692|
      OUT045|                           2002|     Medium|                 Tier 2|Supermarket Type
1|         1133.1916|              907.0|                  1.0|        8.0|                  1.0|
                    2.0|           0.0|                  1.0|           0.0| (1543,[907],[1.
0])|     (4,[1],[1.0])|(15,[8],[1.0])|      (9,[1],[1.0])|            (8,[2],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 803.1797886168849|
|         DRC01|       5.92|          Regular|     0.019308607|Soft Drinks| 49.0692|
      OUT017|                           2007|     Medium|                 Tier 2|Supermarket Type
1|          1478.076|              907.0|                  1.0|        8.0|                  7.0|
                    7.0|           0.0|                  1.0|           0.0| (1543,[907],[1.
0])|     (4,[1],[1.0])|(15,[8],[1.0])|      (9,[7],[1.0])|            (8,[7],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 790.6472831235122|
|         DRC12|      17.85|          Low Fat|     0.037826873|Soft Drinks|189.7188|
      OUT046|                           1997|      Small|                 Tier 1|Supermarket Type
1|         2285.0256|              908.0|                  0.0|        8.0|                  2.0|
                    3.0|           1.0|                  2.0|           0.0| (1543,[908],[1.
0])|     (4,[0],[1.0])|(15,[8],[1.0])|      (9,[2],[1.0])|            (8,[3],[1.0])|(2,
[1],[1.0])|            (2,[],[])|(3,[0],[1.0])|(29,[0,1,5,14,21,...|3067.6173482995546|
|         DRC25|       5.73|          Low Fat|     0.045334098|Soft Drinks| 87.0882|
      OUT013|                           1987|       High|                 Tier 3|Supermarket Type
1|         1803.6522|               79.0|                  0.0|        8.0|                  6.0|
                    6.0|           2.0|                  0.0|           0.0|  (1543,[79],[1.
0])|     (4,[0],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|            (8,[6],[1.0])|
(2,[],[])|         (2,[0],[1.0])|(3,[0],[1.0])|(29,[0,1,5,14,21,...|1342.3755568793158|
|         DRC27|       13.8|          Low Fat|     0.058091482|      Dairy|245.1802|
      OUT035|                           2004|      Small|                 Tier 2|Supermarket Type
1|         5650.6446|             1272.0|                  0.0|        4.0|                  3.0|
                    4.0|           1.0|                  1.0|           0.0|(1543,[1272],[1.
0])|     (4,[0],[1.0])|(15,[4],[1.0])|      (9,[3],[1.0])|            (8,[4],[1.0])|(2,
[1],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,1,5,10,21,...|3999.3506683213764|
|         DRC27|       13.8|          Low Fat|     0.058192802|      Dairy|246.9802|
      OUT049|                           1999|     Medium|                 Tier 1|Supermarket Type
1|         5896.3248|             1272.0|                  0.0|        4.0|                  5.0|
                    5.0|           0.0|                  2.0|           0.0|(1543,[1272],[1.
0])|     (4,[0],[1.0])|(15,[4],[1.0])|      (9,[5],[1.0])|            (8,[5],[1.0])|(2,
[0],[1.0])|            (2,[],[])|(3,[0],[1.0])|(29,[0,1,5,10,21,...|3940.5852931316194|
|         DRC27|       13.8|          Low Fat|     0.058339153|      Dairy|246.2802|
      OUT018|                           2009|     Medium|                 Tier 3|Supermarket Type
2|          1228.401|             1272.0|                  0.0|        4.0|                  0.0|
                    1.0|           0.0|                  0.0|           2.0|(1543,[1272],[1.
0])|     (4,[0],[1.0])|(15,[4],[1.0])|      (9,[0],[1.0])|            (8,[1],[1.0])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[2],[1.0])|(29,[0,1,5,10,21,...|  3612.24217303598|
|         DRC27|       13.8|          Low Fat|     0.097251621|      Dairy|245.7802|
      OUT010|                           1998|     Medium|                 Tier 3|     Grocery Stor
e|          245.6802|             1272.0|                  0.0|        4.0|                  8.0|
                    8.0|           0.0|                  0.0|           1.0|(1543,[1272],[1.
0])|     (4,[0],[1.0])|(15,[4],[1.0])|      (9,[8],[1.0])|                (8,[],[])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[1],[1.0])|(29,[0,1,5,10,21,...|1972.0395500213303|
+----------------+------------+-----------------+---------------+------------+--------+-----
------------+-------------------------+-----------+--------------------+----------------
-+-----------------+----------------+-----------------+----------+-------------------+---
-----------------------+------------+-----------------+------------+----------------
---+-----------------+---------------+-------------------+------------------------+-----
--------+--------------------+-------------+--------------------+------------------+
only showing top 20 rows
```

Evaluate Model

In [ ]:

```
from pyspark.ml.evaluation import RegressionEvaluator
```

In [ ]:

```
errors = ["r2", "rmse", "mse", "mae"]
name = ["R-Square or Accuracy", "Root Mean Square Error", "Mean Square Error", "Mean Abs
```

```
for i in range(len(errors)):
  eval = RegressionEvaluator(predictionCol="prediction", labelCol='Item_Outlet_Sales', m
  print("The {} of Model is {}".format(name[i],eval.evaluate(result)))
```

```
The R-Square or Accuracy of Model is 0.5544914690571232
The Root Mean Square Error of Model is 1159.6628884131828
The Mean Square Error of Model is 1344818.0147628062
The Mean Absolute Error of Model is 853.5910257087111
```

```
In [ ]:
```